Improvements in Spontaneous Speech Recognition

Daben Liu, Long Nguyen, Spyros Matsoukas,
Jason Davenport, Francis Kubala, Rich Schwartz

BBN Technologies, GTE Internetworking
70 Fawcett Street, Cambridge, MA 02138

ABSTRACT

Recognition of spontaneous conversational speech is made difficult by the severe pronunciation variants, mostly due to accents, co-articulation and speaking mode. Most recognition systems only take some of these phenomena into account. The result is that even though they get high accuracy on prepared or carefully read speech, the performance on spontaneous speech is poor. In this paper, we summarize some of the most profound features in spontaneous speech, such as non-speech events, co-articulation, and deleted phonemes. This paper will show that just by introducing some simple but efficient solutions into our BYBLOS recognition system [1], we can significantly improve the performance of recognizing spontaneous speech. The experimental results have shown that for Hub4 96 development test set, the WER for the spontaneous speech (F1) was reduced by about 35% and for the Hub4 96 evaluation set, the improvement was about 25%.

1. INTRODUCTION

Spontaneous speech, as opposed to planned speech, is a more natural way in which people communicate with each other. However, the recognition of spontaneous speech is made more challenging by the severe pronunciation variants and unpredictable pauses or laughter in between words.

Pronunciation variants are largely due to accents, co-articulation, speaking style and/or speaking mode. The variants can be in a word, such as "BECAUSE" which, in fast speech, is usually pronounced as "CUZ". Or they can be in between two words. A common example would be that "GOING TO" is spoken as "GONNA" in most casual conversations. It is observed that, in spontaneous speech, the phonetic realization of many words is quite different from the canonical phonetic pronunciations in our standard dictionary. The most dramatic example is that of phonemes being deleted, which we set as a separate case to study.

Non-speech events, such as filled pauses and laughter, are another major source of confusion for the recognizer. In the BN training transcriptions, we see more than 10,000 filled pauses and in the Hub4 development test set (2 hours), there are more than 500 pauses. Of these, half of them are in the F1 speech, even though F1 accounts for only 30% of the total. It's usually the case that when a pause-filler is incorrectly recognized, the bad influence will extend to the adjacent words. As a result, dealing successfully with these events would be necessary for satisfactory spontaneous-speech recognition performance.

In the following sections, we will try to give an insight to each of the phenomena, one by one, and also present our solutions. Note that, in this paper, "spontaneous" is referring to the "high fidelity spontaneous" speech which is dubbed as "F1" in NIST "Hub-4" annotation specification. In the following sections, if not specified, all the reported results are based on 11-hour male training data chosen from the 80-hour training corpus. Also the test set is on male Dev96 data unless otherwise specified. "Dev96" means NIST Hub-4 1996 development set.

2. PRONUNCIATION VARIANTS AND COMPOUND WORDS

In previous state-of-the-art speech recognition engines, the recognition dictionary is designed such that each word can have more than one pronunciation and the decoder will handle it by allowing alternative paths for the word. In doing this, the pronunciation variants of a single word can be alleviated. However, in fast spontaneous speech, co-articulation between words is also quite common due to some speaking conventions (e.g., GONNA, WANNA…). We have observed that this is one of the major sources that confuse the recognition system. Finke [2] has used the idea of multiword tokens and built 21 rules to model the variability of these tokens, which showed a significant gain.

In our system, since we are already able to handle multiple pronunciations, the procedure we have used is relatively simple and straightforward.

Compound Words in the Dictionary

We started with a short list of 170 compound words that had been used by Finke [2]. We found that it required very little effort to create the alternative pronunciations.

For each of the compound words, a native speaker spoke the words out loud in a casual manner and then typed in the phonetic transcription. This only took about 1 hour for all 170 words. These compound words are then put into the dictionary.

Compound Words in Acoustic and Language Models

All the compound words in the training transcriptions were concatenated and treated as a word in both acoustic training

and language training. The resulting acoustic and language models were then used in the test of Dev96 with BBN Byblos system. Table 1 shows the unadapted test word error rate (WER) results. With just adding the tokens and retraining, the performance on F1 and overall test set has both improved by about 0.5% absolute.

Condition	F1	overall
Without compound words	31.8%	31.4%
With compound words	31.3%	30.9%

Table 1. Performance of system trained with compound words

Phonological Probabilities

After training with multiple pronunciations, we perform a forced HMM alignment on the training data. For each word, we count the number of times that each spelling is chosen in the alignment. These counts are converted to phonological probabilities using the same backoff procedure that we use for our language model [4].

Given a word in the dictionary which has C1 counts for pronunciation 1 (P_1) and C2 for the other (P_2),

if $(C1 == 0 \quad and \quad C2 == 0)$

$p(P\_1) = p(P\_2) = 0.5$

else

if $(C1 == 0 \quad or \quad C2 == 0)$

$\mu = 1$

else

$\mu = 2$

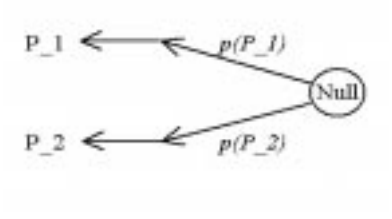
endif

$$p(P\_1) = \frac{C1}{C1 + C2 + \mu} + 0.5 * \frac{\mu}{C1 + C2 + \mu}$$

$$p(P\_2) = \frac{C2}{C1 + C2 + \mu} + 0.5 * \frac{\mu}{C1 + C2 + \mu}$$

endif

Then we apply these phonological probabilities in the backward pass of our 2-pass Nbest decoder [3] and also in the Nbest rescoring. When the word is activated, both pronunciations are activated together, each with its own phonological probabilities, as illustrated below:



Using phonological probabilities improved the performance for another 0.5% absolute, which is shown in Table 2.

Condition	F1	Overall
With compound words, no phonological prob.	31.3%	30.9%
With compound words, with phonological prob.	30.8%	30.5%

Table 2. Results with Phonological Probabilities

Picking the Compound Words

With the 1% absolute gain in hand, we set out to expand the compound word list in a more rational way. To make the new algorithm more efficient, the compound words should be those that occur most frequently. Led by this thought, the 1000 most frequent bi-words were chosen from the LM bi-grams. We used the same informal procedure as with the first 170 compound words. For each word, we decided subjectively whether an alternate pronunciation was needed. We observed that the most frequent bigrams usually had a reasonable alternative. But as we went further down the list, a smaller percentage of words needed alternates. By the end of the list, only about one fifth of the bigrams were assigned alternates. So, given the decreasing probability of the bigrams, and the decreasing probability of alternates, we believe we accounted for the vast majority of compound words that needed alternates. We had a total of 314 compound words in our dictionary. The results with the new list are shown in Table 3. From the results, the total gain on spontaneous speech due to compound word is 2% absolute.

Condition	F1	overall
New compound word list with phonological prob.	29.8%	30.1%

Table 3. Results with 314 compound words

3. PHONEME DELETION AND SKIP TOPOLOGIES

Because spontaneous speech is faster and more casual, many phonemes are deleted or severely shortened. In particular, we have observed that the HMM forced alignment on the broadcast news transcriptions usually results in a lot of minimum phoneme durations and we suspected that the phoneme is actually not spoken by the speaker in many cases. However, we have found that our system generally achieves better accuracy using 5-state phone models than 3-state phone models. To alleviate this problem, we modified the HMM topology for a phoneme so that it allows a shorter transition path, as shown by the dotted line in Figure 1.

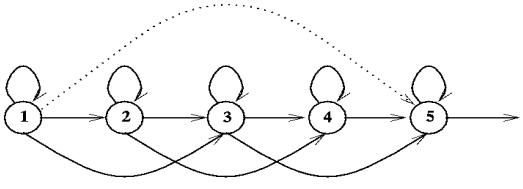


Figure 1 Topology of 5-state HMM with skip transition

To see the effect of the added transition, we trained two non-crossword SCTM 32-bin systems on 11 hours of male and 13

hours of female speech, one using the standard hmm topology and the other using the skip. Then we performed a forced alignment of the references in the 1996 Hub-4 development test (H4D96) using both models and collected the statistics of the three- and two-state phonemes. As we can see in Table 4, the number of 3-frame phoneme occurrences drops dramatically when using the skip hmm topology, which supports our belief that most of the 3-frame phonemes observed with the standard hmm topology are actually shorter, or deleted.

Condition	F1		Overall	
Number of frames	3	2	3	2
Standard 5-state hmm	19.7%	-	14.0%	-
5-state hmm with skip	6.1%	14.5%	5.2%	9.7%

Table 4 Percentages of three- and two-frame phoneme occurrences in H4D96

The skip hmm topology has also significant impact on the recognition word error rate. Table 5 shows the word error rates for both of the above systems on the 1996 Hub-4 UE Development test, where we can see that the skip transition helps improve F1 by 2.0% absolute. As expected, the gain in F1 is larger than the overall gain, since spontaneous speech is by nature faster and more co-articulated.

Condition	F1	Overall
Standard 5-state hmm	37.3%	33.7%
5-state hmm with skip	35.3%	32.8%

Table 5 Recognition results for skip hmm topology

4. NONSPEECH EVENTS

It's usually the case that when people are talking, they make pauses for thinking and organizing their words. In the BN training transcriptions, we see more than 10,000 pause-fillers, such as [UH], [UM], or [HMM], and more than half of them occur in the F1-conditioned speech. The relatively high energy of these pauses makes it difficult for the recognizer to treat them just as silence. With previous system, of the 457 pause-fillers in the Dev96 set, only 110 were correctly recognized. Others are recognized as a word such as "a", "the", "and" or "of" which accounts for a moderate portion of word errors. Besides the pauses, there are other nonspeech events such as laughing, coughing, which may not be as dominant in spontaneous speech but still degrade the system's performance.

To deal with these nonspeech events, we tried some simple adjustments in our system, which turned out to be successful.

Modeling Pauses in Language Model

By explicitly modeling these pause-fillers in the language model, we improved the word error rate. Experiments have shown that the performance, which is listed in Table 6, has been improved significantly.

Condition	F1	Overall
Without pauses in LM	34.6%	33.6%
With pauses in LM	33.0%	32.5%

Table 6 Results with pause-fillers in LM

In the language training transcriptions, we have included a large portion of old corpus (more than 400M words), which don't have any annotations for pause-fillers. Only the new Hub4 training transcriptions with 850k words, that are also included in language training, have the annotations. We augment our LM data with these acoustic training data to explicitly model the pause fillers.

Long Duration Pauses and Modeling Laughter

Some pause-fillers sound very much like real words. For example, the pause-fillers [UH] sounds like the word "A". But pause-fillers are usually quite long. To reduce confusions between pause-fillers and words, we made the dictionary pronunciation of the pause-fillers have 4 phonemes. For example, [UH] is given AH-AH-AH-AH.

Because of the similarity among most pause fillers, we map them into 3 instances, which guarantees that there're enough data for training. Among other nonspeech events, however, it seems that laughter is the only one that has enough training data to make a reasonably good model. As a result, we trained only on the laughter data to build a new model.

After making these modifications and retraining the system, the results on Dev96 are satisfactory, which is shown in Table 7.

Condition	F1	Overall
Without pauses	34.6%	33.6%
With long pauses and laughter	31.8%	31.4%

Table 7 Results with long pauses and laughter

The final improvement from all the nonspeech treatment is 2.8% absolute on spontaneous speech and 1.2% absolute on all conditions.

Missing Targets vs. False Alarms on Pause-Fillers

The word error rates, showed above for the nonspeech events, have drawn a very promising picture. However, taking a closer look on the errors tells us an actual tradeoff in modeling the pause-fillers. When we explicitly model pause-fillers in the LM, we largely increase the precision of recognizing these pauses. At the same time, however, we have increased the possibility of incorrectly recognizing a word as a pause. Table 8 shows the error analysis for the three cases. The first column is the baseline where we do not include pause-fillers in LM. We see no false alarm in this column. The second column is when we include pause-fillers in language training. The minimum duration is only 2 frames long. The missing is less but we have false alarms

now. In the third column, where new models are given to both pause-fillers and laughter, the number of missing is significantly decreased while the false alarms are more than doubled.

Total # of pauses in references: 457

	No pauses	Short pauses	Long pauses with laughter
Total # of pauses in hypothesis	28	229	405
Correctly recognized pauses	110	228	331
Missed pauses	347	229	126
Incorrectly detected pauses (false alarm)	0	33	78
Total errors	347	262	204

Table 8 Analysis of errors caused by pause-fillers.

From the table, we see the false alarms are increasing with each step of progress, though the total number of errors is decreasing. A trade-off is clearly seen and we seem to have achieved a very good balance when coming to the long-duration-pause-filler experiment.

5. TESTS AND COMPARISON

So far, we've achieved improvement in spontaneous speech recognition in different pieces. There still exists the risk of whether these pieces are additive, or if there are negative interference among all these new algorithms and methods.

All the new methods in this paper have been adopted into 1997 Byblos Broadcast New Transcription System and have been tested on both the 96 evaluation test set and 96 development test set. The results show a large improvement on the spontaneous speech (F1). Table 9 shows the improvement on the evaluation set and Table 10 shows the development set, compared with last year's results. All the results are of both genders and unadapted.

Condition	96 system	97 system	Improvement (Relative)
F1	34.2	26.3	23.1%
Overall	34.5	29.0	15.9%

Table 9 Comparison on 96 evaluation set

Condition	96 system	97 system	Improvement (Relative)
F1	39.4	25.4	35.5%
Overall	35.2	26.8	23.9%

Table 10 Comparison on 96 development set

6. CONCLUSIONS

Spontaneous speech is a natural and basic form of day-to-day communication among people. Accurate recognition of this kind of speech is very desirable. However, its naturalness stirs up much more pronunciation variants in spontaneous speech than in read or prepared speech. There are also pause-fillers, laughter and other nonspeech events, all of which make the recognition task difficult.

In this paper, we have studied three different kinds of features in spontaneous speech and proposed the following solutions.

- To alleviate the effect of strong co-articulation between words, we used compound word tokens in our training, estimated phonological probabilities and applied them in the decoder.
- To model the phoneme-deletion phenomenon, we used a new HMM topology allowing a shorter path in our 5-state model.
- To decrease the confusion caused by pause-fillers, laughter or other nonspeech events, we explicitly modeled the pause-fillers in the language model, assigned longer duration for them and trained a new model for laughter.

All of these methods are directly derived from analysis of the physical phenomenon of spontaneous speech and the resulting precision is quite satisfactory. They're very good examples for telling that in dealing with nonstationary speech signal, knowledge of its particular behaviors is very important and helpful.

7 ACKNOWLEDGMENT

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

8 REFERENCE

1 F. Kubala, et al, "The 1997 BBN Byblos Hub-4 Transcription System", *Proceeding of the DARPA Speech Recognition Workshop 1998*, to be seen elsewhere in this proceeding, February 1998, Virginia

2 Mike Finke, Alex Waibel, "Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition", *Proc. Eurospeech'97*, pp. 2379-2382, 1997, Greece

3 L. Nguyen, R. Schwartz, "Efficient 2-Pass N-Best Decoder", *Proceeding of the DARPA Speech Recognition Workshop 1997*, February 1997

4 P. Placeway, et al, "The Estimation of Powerful Language Models from small and Large Corpora", *Proc. ICASSP'93*, pp. II-33-36, April 1993, Minnesota